

D-1 意見分析システムにおける意見抽出方式の検討と評価

Design and Evaluation of an Opinion Extraction Method for the Opinion Analysis System

立石 健二 福島 俊一

Kenji Tateishi and Toshikazu Fukushima

1. はじめに

企業活動において消費者の意見は、新製品開発や商品評価のために有益な情報である。そのため、アンケートに基づくマーケティングリサーチが企業においてしばしば実施されている。しかし、この方法にはアンケート調査自体の実施コストと収集データの分析コストという2つの問題がある。前者は、モニターを選別し、有効な回答を多数集めるには多くの手間が必要であり、特に自由回答欄は選択欄に比べ回答率が低いので収集コストが高いことを意味する。後者は、自由記述文の分析は従来の統計ツールでは困難であり、人手による分析は手間が多かかり、分析過程で見落としや誤りが混入する可能性があることを意味する。

そこで筆者らは、第1の問題に対して、インターネットを情報源として意見を自動抽出するというアプローチを採用した。インターネットは、あらゆる個人が自由に情報を発信できる場であり、そこには他のメディアにはない多くの意見が存在すると期待できる。また、第2の問題に対しては、テキストマイニングの技術を応用し、抽出した意見から主観的分析では困難な隠れた知識の発見するというアプローチを採用した。

筆者らは現在、この情報抽出とテキストマイニングの技術を融合した新しいマーケティングリサーチシステムとして意見分析システムを開発している[1]。このシステムによれば、ユーザが指定した商品が持つ意見の特徴を他の商品群と比較分析することや、その分析を意見の肯定・否定の評価別に実施することが可能である。例えば、「モバイルギアの肯定の意見の特徴を、競合商品「カシオペア」「リブレット」の肯定の意見と比較する」のような分析が可能である。ここで本システムは、Web ページから指定された商品に関する意見を抽出する意見抽出部と、抽出した意見をユーザの条件に従って分析する意見分析部の2つから構成されるが、意見収集部の出力はそのまま意見分析部の入力となるため、分析結果の信頼性は意見抽出部の精度に依存するところが大きい。そこで本稿では、意見抽出部の性能評価を中心に報告し、この意見抽出部を意見分析部の入力として利用する場合の問題点や課題を考察する。

2. 意見抽出方式

まず、意見抽出方式の概要を説明する(詳細は、[2]を参照)。筆者らが定義する意見とは指定された商品に対する評価を示す文章である。評価とは、意見の肯定又は否定の属性である。したがって、意見を抽出するとは、「指定された商品に関する肯定 or 否定の評価を示す文章を抽出すること」を意味する。

この意見を抽出するために筆者らが採用したアプローチは、評価表現辞書を作成することである。これは、意見には必ず物事に対する評価を示す表現(以下、評価表現)が存在するという仮定に基づいたものである。意見の抽出は、この辞書を用いて、指定された商品名と評価表現を含むパッセージのなかで、2つの表現間の意味的関係が強いもの

商品分野	評価表現リスト
共通	好き、いい、良い、勧め、最高、満足だ..
書籍	面白、読みにく、分かりやす、違和感、..
コンピュータ	速い、重い、うるさ、不安定、信頼で、..
食品	美味、不味、絶品、名品、しつこい、..

表1 評価表現辞書の例(斜体は否定の基本評価を、それ以外は肯定の基本評価を意味する)

意見らしさの判定ルール(正規表現)	ルールの意味
商品名_.*評価表現_	商品名が前方
商品名_.*(。 . ?!).*評価表現_	同一文に存在
商品名_({0,12})評価表現_	接近して存在
商品名_({0,12})(は が も).*評価表現_	商品名が主格

表2 意見らしさを判定するルールの例

を抽出することで実現する。また、意見の肯定、否定の評価への分類は、評価表現にあらかじめ肯定又は否定の基本評価を付与しておき、評価表現の近傍に否定表現(「ない」や「非」「不」のような接頭語)が奇数回出現する場合は評価を反転させることで実現する。評価表現辞書の例を表1に示す。辞書には、商品分野毎に分けて評価表現を登録する。

商品名と評価表現の意味的関係の強さ(以下、意見らしさ)を計算する方法については、まず、商品名と評価表現の両者を含むパッセージを意見候補とする。この意見候補に対して、表2のような正規表現のルールを適用し、意見候補が満足するルールの組み合わせに応じたスコアを与える。このスコアはあらかじめサンプルの意見候補から学習によって求めておく。このようにして求めた意見らしさのスコアが閾値を超える意見候補を意見と判定する。

3. 性能評価

2節で説明した方式の有効性を評価するために2つの実験を行った。

3.1 意見抽出精度

筆者らは意見を評価表現と商品名との意味的な関連の強さを計算することで抽出しているが、一方で、意見というカテゴリに文書を分類するテキスト分類の問題として捕らえることもできる。そこで、本節では、2節で説明した意見抽出手法と、単語の頻度を素性空間として用いた機械学習による分類手法の性能を比較する。

コンピュータ関連、アルコール関連という2つの商品分野を対象として、各分野について4つ、計8つの商品名を実験に用いた(PDAを4商品、ビールを4商品)。

実験対象は、これらの商品名をキーワードとして検索エンジンで検索した結果のそれぞれ上位300ページ、計2400ページに含まれるパッセージとした。この内、半分の1200ページ(各商品につき150ページ)に含まれるパッセージを学習用として、残りの1200ページに含まれるパッセージを評価用として使用した。

ここで、本実験におけるパッセージを、商品名を含む1文とその前後の1文(全3文)と定めた。各パッセージに

商品分野	パッセージ	学習データ	評価データ
コンピュータ	2092	1106(201/905)	986(199/787)
アルコール	2016	1083(212/871)	933(194/739)

表3 評価に用いたパッセージ数

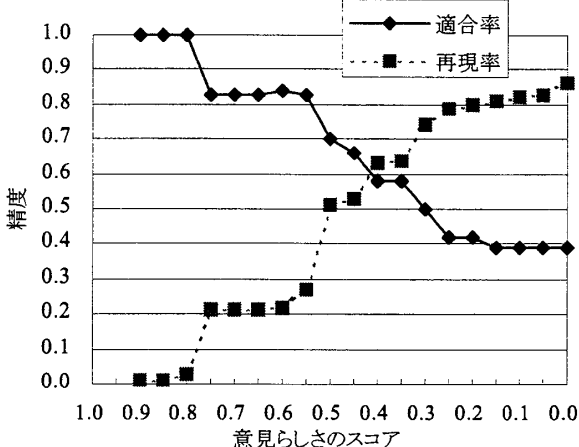


図1 意見抽出精度 (アルコール関連)

は、その商品に対する意見である場合には1、そうでない場合には0のラベル値を付与した。表3に、学習データと評価データのパッセージの数を示す。

次に、1のラベルが付与された学習データから評価表現辞書を作成した。評価表現の数は、コンピュータ関連で132、アルコール関連で63であった。さらに、意見らしさのルールに対するスコアもこの学習データ(先ほどの評価表現を用いて抽出したコンピュータ関連で311、アルコール関連で411の意見候補)から求めた。

一方、比較対象の手法としては、Support Vector Machine(SVM)を用いたテキスト分類手法を選択した。SVMは高性能な学習器の一つとして近年幅広く利用されている。SVMのパッケージとしては、奈良先端科学技術大学院大学の松本研究室で開発されたTinySVM(<http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/>)を使用した。本実験では、SVMの素性として学習データ中のすべての名詞、動詞、形容動詞の出現頻度を用いた。

図1は筆者らの手法でのアルコール関連の抽出精度を示す(コンピュータ関連でもほぼ同様の軌跡を示した)。この図から分かるように適合率は意見らしさのスコアに従って増加する。そのため筆者らの方式で用いた正規表現のルールは商品名と評価表現の関係の強さを適切に計算できたと言える。また抽出結果の上位10件の適合率は、コンピュータ関連では72%、アルコール関連では84%であった。

表4は、筆者らの手法とSVMによるテキスト分類手法を比較した結果である。筆者らの手法において意見らしさのスコアが0.43以上を閾値とした場合の再現率がSVMの再現率と同程度になる。この条件において筆者らの手法はSVM分類よりも高い適合率を持つことが分かる。この結果からテキスト分類の素性として単語の頻度を単純に用いるだけでは十分でなく、筆者らの手法のように商品名と評価表現の関連性を考慮するのが有効と考える。

3.2 意見分類精度

次に、抽出した意見を肯定又は否定の評価へ分類する精度を評価する。4.1節の評価データを対象として筆者らの手法で正解として抽出されたコンピュータ関連で137、アルコール関連で166の意見を実験対象とした。否定表現と

精度	SVM分類	提案手法
適合率	47%(82/172)	62%(86/139)
再現率	41%(82/199)	43%(86/199)

精度	SVM分類	提案手法
適合率	51%(112/219)	59%(118/199)
再現率	57%(112/194)	61%(118/194)

表4 SVM分類と提案手法との比較(上:コンピュータ、下:アルコール)

して「ない」を採用し、否定表現を検出する範囲は評価表現の後方6文字以内とした。

実験結果の適合率はコンピュータ関連で87%(119/137)、アルコール関連で93%(154/166)、全体で90%(273/303)であった。この結果から筆者らの方式は高い分類精度を実現できることがわかる。なお、全303件中、否定表現が出現することにより属性が反転したケースは6件であった。そのため、意見の分類はほぼ評価表現の基本属性のみから求めることが可能である。

4. 考察

3節の結果から、評価表現辞書を用いて意見と評価表現の関連性を考慮した意見抽出方式の妥当性と、この辞書により意見の肯定、否定の分類も非常に容易に実現できることが確認できた。しかしながら、意見抽出部の出力を意見分析部の入力とする場合には、より意見抽出の精度向上が必要と考える。例えば、再現率50%を要件とした場合、現状の適合率はアルコール関連で70%、コンピュータ関連60%程度であり、現状の方式は再現率を求められる場合では精度がやや不十分と考えている。そこで、筆者らの方式で不正解となったケースを分類しその対策を検討する。

- (1) 商品名と評価表現に意味的な繋がりが無い場合
例) コンセプトはNECのモバイルギアとほぼ同じだが、発売されたのが早かったので少々物足りない。
このケースについては、商品名と評価表現の依存関係を構文解析により判定することが必要である。
- (2) 商品のキャッチフレーズや商品紹介が検索される場合
例) 本日のおすすめ商品はモバイルギアです。
このケースについては、あらかじめその商品のメーカーやショッピングサイトを削除する前処理が必要である。
- (3) 商品名が複数の意味を持つ場合
例) ザウルス→PDA or 恐竜
このケースについては、正確には語義の曖昧性の問題であるが、あらかじめ商品分野をキーワードとしてWebページを絞り込んでおくなどの簡単な対策も考えられる。

5. おわりに

本項では、筆者らが現在開発している意見分析システムの意見抽出部の性能評価を中心に報告した。評価結果からは、意見抽出を評価表現辞書を用いて意見と評価表現の関連性を考慮する方式の妥当性が確認できた。今後は、考察で分析した不正解のケースのそれぞれについて対策を講じ、意見分析部の精度向上を目指す。

参考文献

- [1] 立石健二 森永聡 山西健司 福島俊一, "Web上の自動意見分析—情報抽出とテキストマイニングの融合—", 情報処理学会第64回全国大会予稿集(3), pp19-20, 2002.
- [2] 立石健二 石黒義英 福島俊一, "インターネットからの評判情報検索", 情報処理学会研究報告, NL-144-11, pp.75-82, 2